"Diga como me medirá e te digo como me comportarei [te engano]" (Goldratt, E.).

Métricas e Experimentos

Paulo Ricardo Lisboa de Almeida

DSBD

UFPR

Medindo a qualidade

Quando temos um conjunto de dados rotulados, desejamos treinar um modelo **e** avaliar a qualidade desse modelo.

Criar um experimento para verificar como o modelo vai se comportar no mundo real.

Não é tarefa simples e não existem regras pré-definidas para isso.

Depende do problema.

O que existem são técnicas conhecidas. Escolher a técnica e aplicá-la corretamente exige bom senso.

Tarefa de um cientista.

Regra de ouro

Considere um conjunto de dados rotulados X que vamos usar para criar nossos modelos.

Precisamos criar conjuntos de treinamento X_{train} e testes X_{test} a partir de X, de forma que.

$X_{train} \subset X$

$X_{test} \subset X$

$X_{train} \cap X_{test} = \varnothing$

Hold-out

Separar o conjunto de dados X aleatoriamente entre X_{train} e X_{test}, de forma que $X_{train} \cap X_{test} = \varnothing$, e $X_{train} \cup X_{test} = X$.

Acurácia

Dado um conjunto de testes, podemos medir a acurácia do modelo treinado.

A acurácia é dada por:

Acc = instâncias corretas / total de instâncias.

Faça você mesmo #1

Implemente uma função que calcula a acurácia usando o exemplo disponibilizado no Google Colab.

Not So Simple

Considere que temos apenas 5 instâncias da classe Iris Versicolour, enquanto temos 50 instâncias da classe Iris Setosa.

Quais os potenciais problemas com a acurácia?

Not So Simple

Considere que temos apenas 5 instâncias da classe Iris Versicolour, enquanto temos 50 instâncias da classe Iris Setosa.

Quais os potenciais problemas com a acurácia?

Considere que todos os dados foram usados para o teste.

Mesmo um classificador inútil, que classifica todas instâncias como Iris Setosa, tem uma acurácia de:

Acc = 50/55 = 91%

Temos ainda vários outros problemas, como a possibilidade de enviesar os classificadores no treinamento.

Ficará mais claro no futuro.

Acurácia

A acurácia só faz sentido em cenários **balanceados**.

Onde todas as classes no conjunto de testes possuem aproximadamente a mesma quantidade de dados.

Matriz de Confusão

Uma matriz de confusão mostra quantos acertos e erros de classificação foram obtidos pelo classificador para todas as classes.

Útil tanto em cenários balanceados quanto desbalanceados.

Podemos nos informar sobre quais classes estão gerando mais dificuldades para o classificador.

Especialmente útil para problemas com mais de duas classes.

Matriz de Confusão

Considerando um número arbitrário de classes, a matriz tem o seguinte formato:

		Predição			
		Classe 1	Classe 2	Classe 3	...
Classe real (ground-truth)	Classe 1	# itens classificados como classe 1	# itens classificados como classe 2	# itens classificados como classe 3	...
	Classe 2	# itens classificados como classe 1	# itens classificados como classe 2	# itens classificados como classe 3	...
	Classe 3	# itens classificados como classe 1	# itens classificados como classe 2	# itens classificados como classe 3	...
	...	...	...	...	...

Pergunta

Como é a matriz de confusão de um classificador que classificou corretamente todos os dados de teste?

		Predição			
		Classe 1	Classe 2	Classe 3	...
Classe real (ground-truth)	Classe 1	# itens classificados como classe 1	# itens classificados como classe 2	# itens classificados como classe 3	...
	Classe 2	# itens classificados como classe 1	# itens classificados como classe 2	# itens classificados como classe 3	...
	Classe 3	# itens classificados como classe 1	# itens classificados como classe 2	# itens classificados como classe 3	...
	...	...	...	...	...

Exemplo

O que esse resultado diz?

		Predição	
		Iris Setosa	Iris Versicolour
Classe real	Iris Setosa	25	0
	Iris Versicolour	3	0

Exemplo

O que esse resultado diz?

Como calcular a acurácia a partir da matriz de confusão?

		Predição	
		Iris Setosa	Iris Versicolour
Classe real	Iris Setosa	25	0
	Iris Versicolour	3	0

Exemplo

O que esse resultado diz?

Como calcular a acurácia a partir da matriz de confusão?

Acc = soma da diagonal principal / soma dos elementos

Acc = (25 + 0) / (25 + 0 + 3 + 0)= 89%

		Predição	
		Iris Setosa	Iris Versicolour
Classe real	Iris Setosa	25	0
	Iris Versicolour	3	0

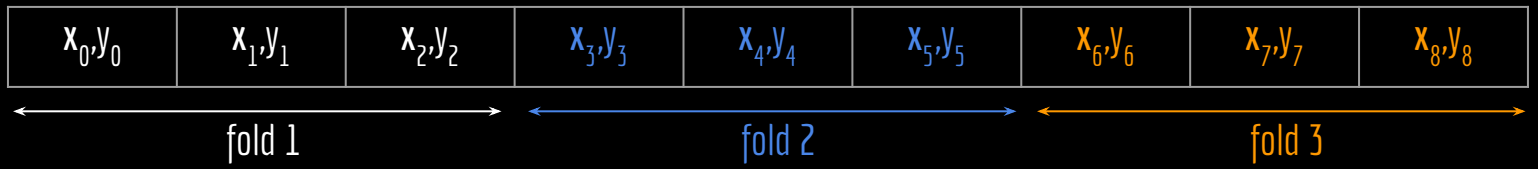
Faça você mesmo #2

Execute o programa disponibilizado algumas vezes, e verifique as matrizes de confusão geradas.

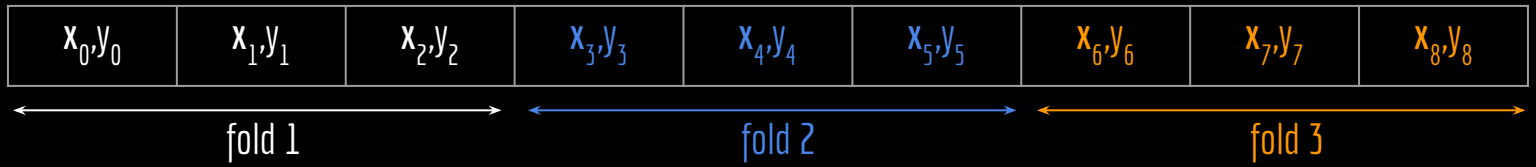
Validação cruzada de k-folds

Quando os dados rotulados são escassos, podemos randomizar os dados, e depois dividi-los em k conjuntos (folds) de tamanhos iguais.

Exemplo. Considere que temos 9 dados rotulados, e k = 3.



Validação cruzada de k-folds




```
Para cada fold f
    Use todos os folds, exceto f, para treinar um modelo
    Teste o modelo em f
```

Validação cruzada de k-folds

Vantagens e desvantagens da validação cruzada de k-folds?

Validação cruzada de k-folds

Vantagens e desvantagens da validação cruzada de k-folds?

+ Todos os dados podem ser usados para os testes, sem sacrificar o tamanho do conjunto de treinamento.
- Pode ser custoso. Precisamos treinar e testar k modelos.
- Precisamos tomar cuidado com as métricas. Por exemplo, como criar uma matriz de confusão que representa a combinação de todos os testes?

Faça você mesmo #3

Verifique como funciona a classe kfold no scikit-learn:

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

Implemente um teste usando k=5 no código disponibilizado.

Mostre a matriz de confusão final, que considera todos os 5 folds. Como você vai fazer isso?

Outras Métricas

Essas foram métricas básicas para problemas de classificação.

Existem várias outras métricas importantes, que podem ser aplicadas, dependendo do seu problema. Pesquise.

- Curva ROC e Área sob a Curva ROC - quando precisamos de um *tradeoff* entre verdadeiros e falsos positivos.
- F1 - para problemas binários onde não nos importamos com verdadeiros negativos.
- Macro F1 - adaptação do F1 para problemas com mais de duas classes.
- Curva Precision Recall e Área sob a Curva Precision Recall.
  Similar ao F1, mas com um tradeoff entre precisão e Recall.
- Acurácia Prequencial - para fluxos contínuos de dados.
- ...

Exercícios

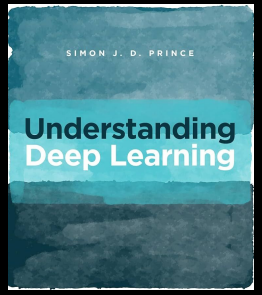
1. Considere que temos relacionados ao clima (umidade, temperatura, pressão, ...) diárias, e um problema onde devemos prever se vai chover ou não amanhã. Os dados estão ordenados no tempo, da seguinte forma:

   dia, umidade, temperatura, ..., classe
   1, 58,25,..., ensolarado
   2, 57,26,..., ensolarado
   3, 100, 20,..., chuva

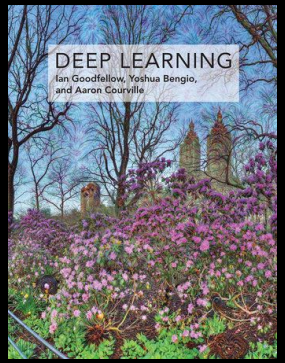
Suponha que você precisa criar um protocolo de treinamento para um perceptron usando esses dados. Nesse caso, ao contrário dos "especialistas" da internet, seria absurdo usar validação cruzada combinada com acurácia, ou similar. Argumente os problemas de usar essa abordagem. Qual seria a sua abordagem para criar um classificador e testá-lo nesse problema?

Referências

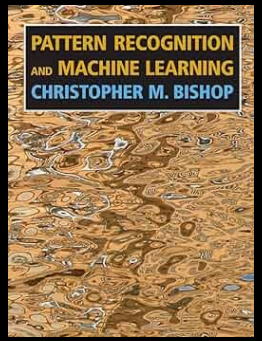
Prince, S. J. Understanding Deep Learning. 2023.



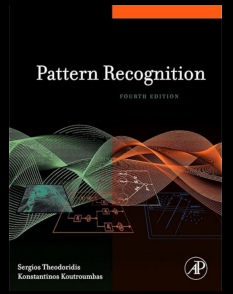
Goodfellow, I., Bengio, Y., Courville, A. Deep Learning. 2016.



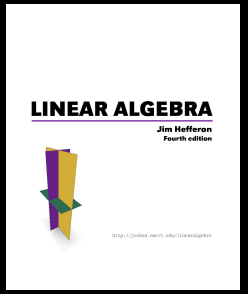
Bishop, C. M. Pattern Recognition and Machine Learning. 2006.



Theodoridis, S., Koutroumbas, K. Pattern Recognition & Matlab Intro. 2010.



Hefferon, J. Linear Algebra. 2015.



Licença